Collegio Carlo Alberto

Model-based clustering using submixtures of Gaussian distributions

Ruth Fuentes-Garcìa

No. 226

December 2011

Carlo Alberto Notebooks

www.carloalberto.org/working_papers

Model-based clustering using submixtures of Gaussian distributions.

Ruth Fuentes-García [1]

December 2011.

Abstract

In a model-based clustering context, one often interested in the estimation of a multivariate mixture model. In this paper we discuss the importance of distinguishing between the number of components in a mixture model and the clusters or groups. We explore the possibility of describing groups as submixtures of Gaussian distributions.

Keywords: Bayesian model-based clustering, Mixture model.

1 Introduction

Several authors have considered clustering problems based on finite mixture models and in particular, mixtures of k Gaussian distributions, see Banfield and Raftery (1993), Bensmail *et al.* (1997). A fully Bayesian approach provides posterior probability information about the number of clusters, component parameters and cluster membership of the observations. In fact, we are interested in the estimation of finite multivariate Gaussian mixtures with an unknown number of components. This not only gives a posterior distribution for the number of components but also provides information on the clustering through the assignment of the observations to the components.

If a mixture of k fixed Gaussian distributions is considered, inference for the model could be carried out through the well known EM or MCMC samplers, such as the Gibbs sampler, (see McLachlan and Peel, 2000). However, if a mixture with an unknown number of components is considered, one possibility is the use of trans-dimensional samplers, see Sisson (2005) for a comprehensive review, to achieve across-model simulations.

In practical applications, the description of one group by only one component of the mixture model may prove to be ambiguous. That is, in general it is inappropriate to assume that a group is well described by a multivariate normal distribution. Hence, recent literature has explored other approaches such as those based on Bayesian nonparametric posterior distributions (see Lijoi *et al.*, 2007), the clustering inherent to the partition induced by Bayesian nonparametric posterior distributions, (see Lau and Green , 2009) or clustering based on the posterior similarity matrix (see Fritsch and Ickstadt, 2009). However, if one thinks of an extreme example given by a dense doughnut shaped cloud of points, it is not clear what a parametric or non-parametric model could conclude in terms of the identification of a single group as would be desirable in some contexts.

In this paper we will consider a feasible model to address the latter situation, a mixture of Gaussian distributions with restricted covariance matrices as the underlying model for the cluster analysis. The fitted model will be used to define groups as

[1]Ruth Fuentes-García is a Lecturer at Facultad de Ciencias, Universidad Nacional Autónoma de México and currently a Visiting Fellow at the Collegio Carlo Alberto, Turin, Italy. *Address for correspondence:* Circuito exterior, Ciudad Universitaria. UNAM. s/n c.p. 04510. México, D.F. E-mail: rfuentes@matematicas.unam.mx. Phone number: +525556223899#45785

sub-mixtures of components, that is each group will be modeled by a subsum of some components of the mixture model. The submixture models will be determined using two criteria in combination. These criteria indicate when a large proportion of observations is swapped between components whilst preventing the merging of components which only swap observations when the sampled parameters are in the tails of the corresponding distributions.

The paper is organized as follows. In section 2 we introduce the basic formulation of the Bayesian hierarchical model emphasizing some difficulties faced during the sampling procedures. Sections 3 and 4 describe the restricted model and the merging criteria respectively. In section 5 we discuss the performance of the method through some examples. Finally in section 6 we give some concluding remarks.

2 Basic formulation in a Bayesian framework

Let $\mathbf{y}_j$ be a random vector, $j = 1, \ldots, n$. There are p measurements of the features of interest for the j-th observation. The density $f(\mathbf{y}_j|\Psi)$ is assumed to be a k-component finite mixture of the form

$$f(\mathbf{y}_j|\Psi) = \sum_{i=1}^{k} w_i \, \mathrm{N}_p(\mathbf{y}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{2.1}$$

where N_p denotes a multivariate normal density in $\mathbb{R}^p$ and
$\Psi = (\mathbf{w}, (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \ldots, (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))$, with $\mathbf{w} = (w_1, w_2, \ldots w_{k-1})$, $0 \leq w_i \leq 1$ and $\sum_{i=1}^{k} w_i = 1$.

From the point of view of cluster analysis, we are mainly interested in identifying and interpreting this underlying mixture structure and its relationship with the sample observations. In order to do so we make use of allocation variables which indicate the cluster membership of each observation.

Consider a vector of categorical random variables Z_j that take values in $1, 2, \ldots, k$. Regarded as allocation variables for the observations, they are assumed to be independent draws from the distributions

$$pr(Z_j = i|\Psi) = w_i \qquad \text{for} \quad i = 1, 2, \ldots, k.$$

Conditional on the $Z_j = i$, the density of $\mathbf{Y}_j$ is given by $f(\mathbf{y}_j|\boldsymbol{\theta}_i)$. If the allocation probabilities are of interest, they are given as

$$pr(Z_j = i|\mathbf{y}_j, \Psi) = \frac{w_i \mathrm{N}_{p_i}(\mathbf{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^{k} w_l \mathrm{N}_{p_l}(\mathbf{y}_j; \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l))}, \quad \text{for} \quad i = 1, \ldots, k.$$

The vector $(z_1, z_2, \ldots, z_n)^T$ is frequently called the *missing data* part of the sample. In many schemes Z is integrated out, but for clustering it plays a pivotal role.

Assume k, is unknown and modeled by a prior distribution with density $p(\cdot)$. The unknown parameters $\Psi = (k, \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}))$ are drawn from a set of appropriate prior distributions. Suppose that a joint prior for $\Psi = (k, \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\Sigma}))$ is given for each k in a countable set $\mathcal{K}$. Here it is assumed that all probability densities are proper. Considering the allocation variables and imposing further conditional independence that

reflects the fact that allocation variables summarize information on the number of components and its weights, so that $p(\boldsymbol{\mu}, \boldsymbol{\Sigma})|\mathbf{z}, \mathbf{w}, k) = p(\boldsymbol{\mu}, \boldsymbol{\Sigma})|k)$ and the likelihood $p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}, k) = p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{z})$, then

$$p(k, \mathbf{w}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{y}) = p(k)p(\mathbf{w}|k)p(\mathbf{z}|\mathbf{w}, k)p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|k)p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{z}). \tag{2.2}$$

where $p(\cdot|\cdot)$ is used to denote generic conditional distributions. We want to sample from $p(k, \mathbf{w}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{y})$, which is proportional to $p(k, \mathbf{w}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{y})$. Several alternatives to sample from the latter are given in recent literature. These methods are extensions of MCMC methods and known as trans-dimensional MCMC. One common choice is the reversible jump MCMC proposed by Green (1995), with multivariate extensions discussed for example in Dellaportas and Papageorgiou (2006), Fuentes-García (2004) and Zhihua *et al.* (2004). Another option is the BDMCMC proposed by Stephens (2000 A), this method is used to estimate the models discussed in this paper. The BDMCMC sampler consists of the construction of a continuous time Markov birth-death process with the appropriate stationary distribution.

Let Ω_k denote the parameter space of the mixture model with k components. Ignoring the labelling of the components, let $\Omega = \cup_{k\geq 1}\Omega_k$. In a k component mixture model the set of exchangeable parameters $\{(w_1, \boldsymbol{\theta}_1), \ldots, (w_k, \boldsymbol{\theta}_k)\}$ can be seen as points in $[0, 1] \times \Theta$ where $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ and $\sum_{i=1}^{k} w_i = 1$.

Suppose that the prior distribution for $(k, \mathbf{w}, \boldsymbol{\theta})$ given the corresponding hyperparameters denoted as ω is of the form $r(k, \mathbf{w}, \boldsymbol{\theta}) = p(k|\omega)p(\mathbf{w}, \boldsymbol{\theta}|k, \omega)$. The posterior distribution $p(k, \mathbf{w}, \boldsymbol{\theta}|\mathbf{y}, \omega)$ is then seen as a marked point process on $[0, 1] \times \Theta$, with each $\boldsymbol{\theta}_i$ associated with a mark $w_i \in [0, 1]$. The number of components is allowed to vary by continuous births and deaths of new model parameters defined by a continuous time Markov process with $p(k, \mathbf{w}, \boldsymbol{\theta}|\mathbf{y}, \omega)$ as stationary distribution keeping ω fixed. This process combined with standard MCMC update steps creates a Markov chain with stationary distribution $p(k, \mathbf{w}, \boldsymbol{\theta}, \omega|\mathbf{y})$.

Consider a k component mixture where parameters are denoted y=$\{$(w_1,$\boldsymbol{\theta}_1$),. . ., (w_k,$\boldsymbol{\theta}_k$)$\}$. Births and deaths are restricted so that when the process is at y at time t, if a birth occurs, the process jumps to a new state with $k + 1$ components:

$$y \cup (w, \boldsymbol{\theta}) := \{(w_1(1 - w), \boldsymbol{\theta}_1), \ldots, (w_k(1 - w), \boldsymbol{\theta}_k), (w, \boldsymbol{\theta})\},$$

if a death ocurrs the process jumps to a new state, with $k - 1$ components:

$$y \setminus (w_i, \boldsymbol{\theta}_i) := \left\{ \left(\frac{w_1}{(1 - w_i)}, \boldsymbol{\theta}_1 \right), \ldots, \left(\frac{w_k}{(1 - w_i)}, \boldsymbol{\theta}_k \right) \right\}.$$

When the process is at y, births and deaths occur as independent Poisson processes. Births occur with overall rate $\beta(y)$ and new parameters are chosen according to density $b(y; (w, \boldsymbol{\theta}))$. A component dies independently of others with rate

$$\delta_j = d(y \setminus (w_j, \boldsymbol{\theta}_j); (w_j, \boldsymbol{\theta}_j)),$$

for some $d : \Omega \times ([0, 1] \times \Theta) \to \mathbb{R}^+$, the overall death rate is given by $\delta(y) = \sum_j \delta_j(y)$. The time to the next birth/death event is then exponentially distributed with mean $1/(\beta(y) + \delta(y))$ and it will be a birth with probability

$$Pr(\text{birth}) = \frac{\beta(y)}{(\beta(y) + \delta(y))}$$

and a death of component j with probability

$$pr(\text{death}) = \frac{\delta_j(y)}{(\beta(y) + \delta(y))}.$$

Stephens (2000 A) showed that, assuming the general hierarchical prior on $(k, \mathbf{w}, \boldsymbol{\theta})$ given by $r(y) = p(k|\omega)\tilde{p}(\boldsymbol{\theta}_1|\omega)\tilde{p}(\boldsymbol{\theta}_2|\omega)\cdots\tilde{p}(\boldsymbol{\theta}_k|\omega)$ and keeping ω fixed, the birth and death process defined above has the desired stationary distribution, provided that when the process is at $y = \{(w_1, \boldsymbol{\theta}_1), \ldots, (w_k, \boldsymbol{\theta}_k)\}$, each point $(w_j, \boldsymbol{\theta}_j)$ dies independently of the others as a Poisson process with a rate

$$d(y \setminus (w_j, \boldsymbol{\theta}_j); (w_j, \boldsymbol{\theta}_j)) = \lambda_b \frac{L(y \setminus (w_j, \boldsymbol{\theta}_j))}{L(y)} \frac{p(k-1|\omega)}{kp(k|\omega)}, \tag{2.3}$$

where $L(y)$ is the likelihood in state y and

$$\begin{aligned} \beta(y) &= \lambda_b, \quad \text{a constant,} \\ b(y, (w, \boldsymbol{\theta})) &= k(1-w)^{k-1}\tilde{p}(\boldsymbol{\theta}|\omega). \end{aligned}$$

for $\tilde{p}(\boldsymbol{\theta}|\omega)$ taken as the sampling distributions for the birth move in the reversible jump sampler. The simulation of the process involves the following steps, starting with the initial model $y = \{(w_1, \boldsymbol{\theta}_1), \ldots, (w_k, \boldsymbol{\theta}_k)\}$:

1. Compute the death rate for each component given by equation (2.3).
2. Compute the total death rate $\delta(y) = \sum_j \delta_j(y)$.
3. Simulate the time to the next jump from an exponential distribution.
4. Simulate the type of jump and adjust y to reflect the jump.
5. Return to step 1.

Assuming the necessary conjugate priors for the hierarchical model, the algorithm described above is combined with MCMC update steps allowing ω to vary. To do this we use Gibbs sampler steps which consider the allocation variables. Given the state $\Phi^{(t)} = \phi^{(t)}$ at time t, simulate a value for $\Phi^{(t+1)} = \phi^{(t+1)}$ by

1. Sampling $(k^{(t)'}, \mathbf{w}^{(t)'}, \boldsymbol{\theta}^{(t)'})$ by running the birth and death process for a fixed time t_0, starting from $(k^{(t)}, \mathbf{w}^{(t)}, \boldsymbol{\theta}^{(t)})$ and fixing ω to be $\omega^{(t)}$. Set $k^{(t+1)} = k(t)'$.
2. Sample $z_j^{(t+1)}$ from $p(z_j = i|k^{(t+1)}, \mathbf{w}^{(t)'}, \boldsymbol{\theta}^{(t)'}, \omega^{(t)}, \mathbf{y})$.
3. Sample $\omega^{(t+1)}$ from $p(\omega|k^{(t+1)}, \mathbf{w}^{(t)'}, \boldsymbol{\theta}^{(t)'}, \mathbf{y}, \mathbf{z})$.
4. Sample $\mathbf{w}^{(t+1)}, \boldsymbol{\theta}^{(t+1)}$ from $p(\mathbf{w}, \boldsymbol{\theta}|k^{(t+1)}, \omega^{(t+1)}, \mathbf{y}, \mathbf{z})$.

In a mixture model, if all the k components belong to the same parametric family, then $f(\mathbf{y}|\Psi)$ is invariant under the $k!$ permutations of the component labels in Ψ. This is known as *label switching* and causes identifiability problems, for example when following the evolution history of components during an MCMC sampler. This problem is often

handled by imposing an artificial *identifiability constraint* on Ψ, for example ordering the mixing proportions so that $w_1 \leq w_2 \leq \ldots \leq w_k$. However, several authors (Celeux *et al.* (2000), Richardson and Green (1997), Stephens (2000 B) and Frühwirth-Schnatter (2001)) have pointed out that this does not always give a satisfactory solution. That is, the constraint may not be effective in breaking the symmetry of the prior, leaving densities that exhibit multimodality. Although our main interest is not to estimate the density but to obtain information on the number and composition of groups in the observed population, label switching must be taken into account at the simulation stage as it is a prerequisite for convergence of a trans-dimensional sampler, see Jasra *et al.* (2005).

In particular, following Stephens (2000 B), a simple procedure was used in Celeux *et al.* (2000) to identify one modal region and estimate the component parameters. We used this procedure and in particular we concentrate on the component means. The method selects one modal region using the early iterations of the Markov chain Monte Carlo sampler. The choice of the initial set of iterations is not highly sensitive but should be enough to ensure that the resulting estimates are a reasonable approximation of the posterior means and should correspond to a set observed before the label switching occurs. The component labels corresponding to the following iterations are permuted according to a k-means-type algorithm to select the permutation that is closest to the current set of means as described below.

Following this procedure, we post-processed the output of the BDMCMC sampler. We consider the sequence of d-dimensional vector samples of size T, conditional on the number of components k, $\psi^1, \ldots, \psi^m$, where $d = kp$, $\psi^i = (\mu_{1,i}, \ldots, \mu_{p,i}, \ldots, \mu_{k,i}, \ldots, \mu_{k,p})$ and m is the longest period observed before the label switching occurs. We focused on the mean values because they have shown to be the most stable parameter which is covered rapidly by the sampler. The length of period m is determined in practice by looking at the number of observations allocated to each component. The period m is the period before the number of components allocated to each component has changed in more that one observation for all components.

Initial reference centers for $j = 1, \ldots, d$ are defined as

$$\overline{\psi}_j = \frac{1}{m} \sum_{i=1}^{m} \psi_j^i,$$

with corresponding variances

$$s_i = \frac{1}{m} \sum_{i=1}^{m} (\psi_j^i - \overline{\psi}_j)^2,$$

We denote $s_i^{[0]} = s_i$ for $i = 1, \ldots, d$. If we take $\overline{\psi}^{[0]} = \overline{\psi}$, the other $(k-1)$ centers can be deduced by permuting the labelling of the mixture components. The rth iteration is relabelled with the permutation j that minimizes the normalized square distance,

$$\|\psi_j^{m+r} - \overline{\psi}^{[r-1]}\| = \sum_{i=1}^{d} \frac{(\psi_j^{m+r} - \overline{\psi}_i^{[r-1]})^2}{s_i^{[r-1]}},$$

for $j = 1, \ldots, k$ and $r = 1, \ldots, T - m$.

The centers and normalizing coefficients are updated after each iteration so

$$\begin{aligned}\overline{\psi}^{[r]} &= \frac{m+r-1}{m+r}\,\overline{\psi}^{[r-1]} + \frac{1}{m+r}\,\psi^{m+r} \\ s_i^{[r]} &= \frac{m+r-1}{m+r}\,s_i^{[r-1]} + \frac{m+r-1}{m+r}\,(\overline{\psi}_i^{[r-1]} - \overline{\psi}_i^{[r]})^2 + \frac{1}{m+r}\,(\psi_i^{m+r} - \overline{\psi}_i^{[r]})^2.\end{aligned}$$

We have found the procedure to be efficient and helpful, it allows us to identify posterior component parameters for the fitted mixture so that their values can be used to determine the submixture of components that describes a group.

The assessment of convergence in trans-dimensional MCMC samplers is particularly difficult. As Brooks *et al.* (2003 B), Castelloe and Zimmerman (2004) and Sisson and Fan (2007) highlight, the challenge lies in finding parameters that retain the same interpretation throughout different models.

We consider the proposal Castelloe and Zimmerman (2004), which is viable for high dimensional problems, to assess convergence for the BDMCMC sampler. Generally speaking, we will select a group of observations from the data set and monitor the parameters of the components to which these data are allocated at each iteration, denoted $\boldsymbol{\theta}_*$. Using the number of components k in the current iteration as the parameter that retains the same interpretation across models. We run $C > 1$ chains of the trans-dimensional MCMC sampler, which are started from overdispersed states with the same number of sweeps. Since the monitored observations are allocated at the end of each sweep, this would allow us to overcome the label switching problem. This set of parameters retains a coherent interpretation across models, a crucial feature for the convergence assessment of trans-dimensional samplers. The selected observations are chosen so that their behavior is expected to vary across sweeps of the sampler in different ways. Namely, we will look for data that are between two clusters, that is between potential competitors when allocating the observations, potential outliers and also data near the centre of a cluster. The convergence assessment looks for evidence that indicates lack of convergence for the set of monitored parameters both across iterations and across chains.

Then, a number m of successive overlapping *batches* for each chain are analyzed. The length of these batches increases and each length is a multiple of a base batch length b. The convergence diagnostic aims to find conditions that would indicate that convergence has not been reached. Some aspects it will detect are: variation between chains; an interaction between models and chains, which indicates between-model variation that differs from chain to chain; and significant differences in frequencies of model visits from one chain to another. The convergence diagnostics are based on the following quantities, which could be interpreted as: $\widehat{V}$, the total variation; Wc, variation within chains; Wm variation within models and $WmWc$, variation within models and chains. They showed that the ratio $\mathbf{E}\widehat{V}/\mathbf{E}Wc \geq 1$, with $\mathbf{E}\widehat{V}/\mathbf{E}Wc = 1$ indicating the absence of a chain effect. The greater the value of this ratio, the stronger the chain effect. Both, numerator and denominator, stabilize as $T \to \infty$. It can also be shown that the ratio $\mathbf{E}Wm/\mathbf{E}WmWc \geq 1$, with $\mathbf{E}Wm/\mathbf{E}WmWc = 1$ indicating: (a) the absence of chain effect, (b) the absence of chain$\times$model interaction and (c) either no model effect or equality of the set of within-chain model frequencies across chains or both. The greater the violation of any of these aspects, the larger this ratio becomes. Here the authors emphasized that the sensitivity of this ratio to the violation of the mentioned aspects is not yet fully understood in terms of the relative weight of the three aspects as $T \to \infty$.

Therefore, the convergence diagnostics proposed which are based on both ratios, are called *potential scale reduction factors (PSRF)*. Using these ratios any potential violations of convergence is monitored. For a parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$, we have

$$PSRF1(\theta_i) = \frac{\mathbf{E}\widehat{V}(\theta_i)}{\mathbf{E}Wc(\theta_i)}, \tag{2.4}$$

$$PSRF2(\theta_i) = \frac{\mathbf{E}Wm(\theta_i)}{\mathbf{E}WmWc(\theta_i)}. \tag{2.5}$$

A multivariate version is also defined to monitor the entire vector rather than considering each element separately. The corresponding multivariate scale reduction factors are

$$MPSRF1(\boldsymbol{\theta}_*) = \text{maximum eigenvalue of } [Wc(\boldsymbol{\theta}_*)]^{-1}\widehat{V}(\boldsymbol{\theta}_*), \tag{2.6}$$

$$MPSRF2(\boldsymbol{\theta}_*) = \text{maximum eigenvalue of } [WmWc(\boldsymbol{\theta}_*)]^{-1}Wm(\boldsymbol{\theta}_*). \tag{2.7}$$

We simulate $C = 3$ chains of equal length T with overdispersed starting values and choose a base batch size b, Brooks *et al.* (2003 B) suggested for example $b \approx \frac{T}{20}$. For $q = 1, \ldots, \frac{T}{20}$, we compute $PSRF1^{(q)}(\theta_i)$, $PSRF2^{(q)}(\theta_i)$, $MPSRF1^{(q)}(\boldsymbol{\theta}_*)$ and $MPSRF2^{(q)}(\boldsymbol{\theta}_*)$, the latter two when numerically available. We look for the q_0 such that for $q > q_0$: (a) the plots for $PSRF1^{(q)}(\theta_i)$, $PSRF2^{(q)}(\theta_i)$, $MPSRF1^{(q)}(\boldsymbol{\theta}_*)$ and $MPSRF2^{(q)}(\boldsymbol{\theta}_*)$ are close to 1; (b) the plots for pairs of numerator and denominator for $PSRF1^{(q)}(\theta_i)$, $PSRF2^{(q)}(\theta_i)$, maximum eigenvalue of $Wm(\boldsymbol{\theta}_*)$ and maximum eigenvalue of $WmWc(\boldsymbol{\theta}_*)$ have settled approximately to a common value. The first $q_0 b$ observations could then be discarded and use the remaining ones used for inference.

3 Restricted covariance mixture model

We consider a mixture of spherical multivariate normal distributions with a probability density function which has a simplified covariance structure, given by the following equation:

$$f(\mathbf{y}_j; \Psi) = \sum_{i=1}^{k} w_i \mathrm{N}_p(\mathbf{y}_j | \boldsymbol{\mu}_i, \tau_i^{-1} I_p), \tag{3.1}$$

where I_p denotes the p dimensional identity matrix.

The importance of allowing the volumes of the normal components to be different when considering the same shape and the same orientation has been pointed out by Celeux and Govaert (1995). They showed that these models are capable of detecting many clustering structures without needing complex algorithms. However, they only considered two dimensional data. Our main concern is to prevent the use of highly dispersed distributions misrepresenting the data. We allow the τ's to vary from component to component to preserve some flexibility in the model but place a tight restriction on their size through the prior distributions we assign, as we shall now describe.

Consider observations $\mathbf{y}_1, \ldots, \mathbf{y}_n$, where $\mathbf{y}_j$ has a distribution given in equation (3.1).

The Bayesian hierarchical model has the following prior structure

$$\begin{aligned} w &\sim \text{Dirichlet}(\delta_1, \ldots, \delta_k), \\ \boldsymbol{\mu}_i &\sim \text{N}_p(\xi, \kappa^{-1}), \\ \tau_i &\sim \text{Gamma}(\alpha, \alpha\beta), \end{aligned}$$

for $i = 1, \ldots, k$. Where ξ is an $p \times 1$ vector, κ is a $p \times p$ matrix and δ_i, α and β are scalars. Note that $\mathbf{E}[\tau_i] = 1/\beta$.

The corresponding posterior full conditional distributions for the Gibbs sampler step in the trans-dimensional algorithm that will be used are

$$\begin{aligned} w &\sim \text{Dirichlet}(\delta_1 + n_1, \ldots, \delta_k + n_k), \\ \boldsymbol{\mu}_i|\cdots &\sim N((\kappa + n_i\tau_i I_p)^{-1}(n_i\tau_i I_p \overline{\mathbf{y}}_i + \kappa\xi), (\kappa + n_i\tau_i I_p)^{-1}), \\ \tau_i|\cdots &\sim \text{Gamma}(\alpha + (pn_i)/2, \alpha\beta + \frac{1}{2}\sum_{j=1}^{n_i}(\mathbf{y}_j - \boldsymbol{\mu}_i)' I_p (\mathbf{y}_j - \boldsymbol{\mu}_i)), \end{aligned}$$

where $n_i = \#\{j : z_j = i\}$ for the allocation variables z_j and $\overline{\mathbf{y}}_i = 1/n_i \sum_{\{j:z_j=i\}}^{n_i} \mathbf{y}_j$.

In particular, we take ξ_j as the midpoint of the corresponding observed interval of variation. Let R_j denote the length of these intervals for $j = 1, \ldots, p$, the matrix κ is a diagonal matrix given as

$$\kappa = \begin{pmatrix} 1/R_1^2 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & 1/R_p^2 \end{pmatrix}.$$

The values for the scalars are taken as follows: $\delta_i = 1$, $\alpha = R_{max}^2$ and $\beta = \sqrt(\gamma)/R_{max}$, where $R_{max} = \max\{R_1, \ldots, R_p\}$. We wish to make a sensible choice of the parameters α and β to ensure that we induce the use of more components without heading for the extreme case where a lot of observations are isolated.

To determine the value of γ we consider an initial analysis of the data set. This constant will give information on whether the groups in the data set exhibit important gaps or they are likely to overlap. We consider the projection of all data points onto the direction of the variable with the maximum observed range R_{max}. The projected data $(y_1', y_2', \ldots, y_n')$ are used to find the largest gap between adjacent y_i''s, defining γ as the largest difference between adjacent y_i''s. When there are well separated groups in the data set, γ tends to be much larger than when the groups overlap. We do not claim these values to be optimal, however, in terms of an exploratory technique they offer a option to obtain initial clustering results.

4 The Merging Criteria

We are interested in criteria to indicate which of the fitted components might be merged to form a submixture that represents the same group. We propose two criteria: one will consider the proportion of allocated data that are *swapped* between components

throughout the sampler and the other will give information on the distance between the densities of the components based on the *affinity* .

Broadly speaking, we are interested in measuring the distance between components. If a group is described by more than one component we expect their distributions to be " close ". Now, when the distance between two components is small, we want to learn about the proportion of observations whose ownership is disputed by the two components to which they are allocated. If this proportion is large, then the components are more likely to be describing one cluster. We suggest that a possible way to obtain straightforward information on the proportion of disputed observations is simply to look at how much data is swapped between components from iteration to iteration. We suggest to merge two components into one group if the proportion of observations exchanged between these components is large and the distance between the components is small, assessed by the affinity between two components. Using these criteria in combination gives information on particular situations. First it helps detecting when a pair of components swaps observations when values on the tails of the distributions of the component parameters are sampled.

Computations to obtain the proportion of swapped observations and the affinity between pairs of component will be made conditional on a given value of k, the number of components. The output of the trans-dimensional sampler will be post-processed. We expect results based on different number of components, which have similar and high posterior probability, to be consistent in terms of the groups they define. Once a value of k is fixed, the label switching is removed, identifying values for the parameters of each component and corresponding allocation vector for the data at each iteration.

4.1 Proportion of observations swapped between components

For the computation of the swaps, consider the allocation variable $Z_{i,j}$ for observation $i = 1, \ldots, n$ at iteration $j = 1, \ldots, J$, which takes a value $k = 1, \ldots, k_j$. To deal with the label switching we have extracted all the cases for $k_j = k$, that is, the calculation is done with an output that includes all the variables for a fixed value of k. Suppose that there are $R \leq J$ of such iterations.

We then consider the allocation variable Z_{ir} for observation $i = 1, \ldots, n$ at iteration $r = 1, \ldots, R$, which takes a value in $(1, \ldots, k)$. We compare it with the value for the same observation i but in the next iteration $r + 1$. A $k \times k$ matrix C is initialized to zeros and for $h = 1, \ldots, k$ and $l = h + 1, \ldots, k$, we add $1/(n_{h,r} + n_{l,r})$ to the element $C_{(h,l)}$, if either $Z_{i,(r+1)} = h$ and $Z_{i,r} = l$ or $Z_{i,(r+1)} = l$ and $Z_{i,r} = h$, where $n_{h,l}$ and $n_{l,r}$ are the number of observations allocated to components h and l at iteration r. Once the matrix is computed for all r, we divide it by R to get an approximate value of the proportion of observations that were exchanged between components.

4.2 The affinity between components

The Bhattacharyya (1943) distance also known as the affinity is used as a measure of similarity between two probability distributions. The affinity between distributions P_1

and P_2 with corresponding densities p_1 and p_2 is defined as

$$A(p_1, p_2) = \int_{\mathbf{y}} \sqrt{p_1}\sqrt{p_2}\, d\mathbf{y}$$

The affinity is related to the Hellinger distance between two distributions P_1 and P_2, with corresponding densities p_1, p_2 which is given as

$$H(P_1, P_2) = \left\{ \int_{\mathbf{y}} (\sqrt{p_1} - \sqrt{p_2}\,)^2\, d\mathbf{y} \right\}^{\frac{1}{2}} \tag{4.1}$$

Consider now H^2

$$H(P_1, P_2)^2 = 2 - 2\int_{\mathbf{y}} \sqrt{p_1}\sqrt{p_2}\, d\mathbf{y} \tag{4.2}$$

Hence, $H = \sqrt{2 - 2A(p_1, p_2)}$, where $A(p_1, p_2)$ is the affinity between P_1 and P_2.

The affinity between two multivariate normal distributions, $N(\boldsymbol{\mu}_1, \Sigma_1)$ and $N(\boldsymbol{\mu}_2, \Sigma_2)$ is obtained in the Appendix.

In the particular case where Σ_1 and Σ_2 are diagonal matrices, the affinity between p_1 and p_2 is given by

$$A(p_1, p_2) = \prod_{i=1}^{p} \left(\frac{2\sigma_{1,i}\sigma_{2,i}}{\sigma_{1,i}^2 + \sigma_{2,i}^2} \right)^{\frac{1}{2}} \exp\left\{ -\frac{1}{4} \sum_{i=1}^{p} \frac{(\boldsymbol{\mu}_{1,i} - \boldsymbol{\mu}_{2,i})^2}{\sigma_{1,i}^2 + \sigma_{2,i}^2} \right\}.$$

We compute the average affinity between all pairs of components. We denote as A the $k \times k$ matrix that shows the average value of the affinity for each pair of the k components, we will refer to it as the average affinity matrix. This will be done for all iterations of the sampler where the specified value of k is observed.

5 Examples.

We will consider two data sets often used in literature which discusses fuzzy and robust clustering to make an assessment of the performance of the proposed methodology.

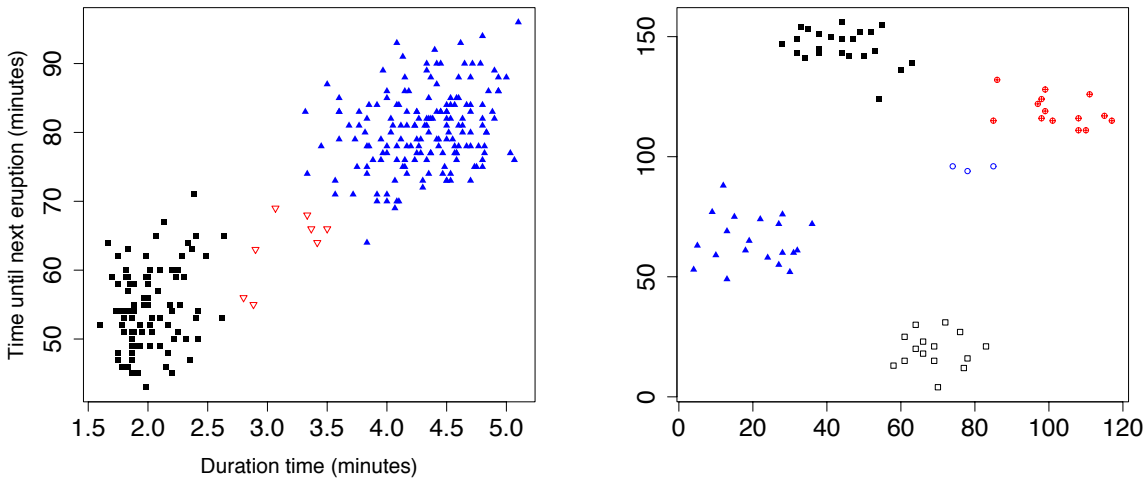
Old Faithful data

This well known data set comprises the duration (mins) and waiting time (mins) before the next eruption for 272 eruptions of the Old Faithful geyser in Yellowstone National Park (the data version used in Härdle (1991), Venables and Ripley (1997) and Stephens (2000 A)),

Ruspini data

As a second example we consider the well known artificial Ruspini (1969) data set. The fact that small weighted components often reflect departures from normality becomes evident with this data set.

Before we discuss the results obtained using submixtures to describe clusters we display the above mentioned data sets, with the resulting classification we would obtain by fitting a mixture of Gaussian distributions with unrestricted covariance matrices and an unknown number of components through a BDMCMC sampler as described in Stephens (2000 A). We select the model defined by the posterior mode for the number of components k, namely $p(k = 3) = 0.66$ for the Old Faithful dataset and $p(k = 5) = 0.47$ for the Ruspini dataset. The classification for all the models we will be discussing is obtained following O'Hagan in the discussion of Richardson and Green (1997). He suggested estimating the clusters by generic hierarchical clustering with dissimilarity based on the number of times each pair of observations occurs together in the same component. Figure 1(a) shows the corresponding three component and five component estimate obtained by this approach with average linkage aggregation, the results were robust to the choice of aggregation method. It is worth mentioning that it has been widely discussed in literature that the posterior distribution for the number of components k is highly dependent on the prior assumptions taken.

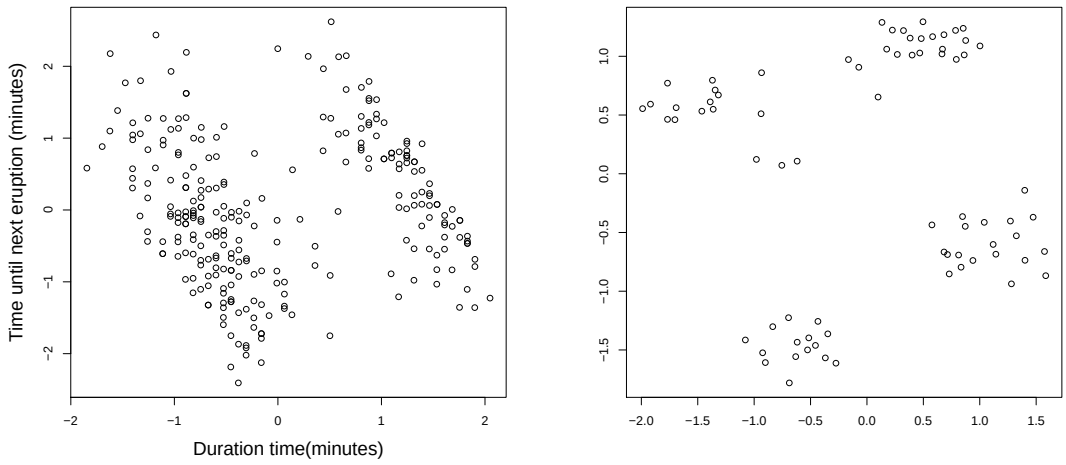


(a) (b)

Figure 1: Classification for the unrestricted model: (a) Old Faithful data; (b) Ruspini data.

Results

The primary task in clustering data is to establish the number and composition of the clusters. In general terms, it would be desirable to have a way of assessing whether the smaller components described in Figure 1 are a different group or are used to cope with departures from normality. We will consider the two data sets we have described rescaled so that they are centered at zero and they have the identity matrix as covariance matrix, see Figure 2. This will be more in accordance with the restricted model if there are no groups in the population of interest. The MCMC sampler described in section 2 was used to fit the mixtures of restricted multivariate normal distributions to two data sets. In each case the samplers were run for a burn-in period of 300000 iterations followed by 100000 iterations, thinned every 50, the resulting 2000 iterations were used for inference.



(a) (b)

Figure 2: Transformed data sets: (a) Old Faithful data; (b) Ruspini data.

The prior assumptions are kept as described in section 2. Values for t_0, the fixed time for which the birth and death process is run, and λ_b, the overall birth rate, are required to use the algorithm. We will consider $t_0 = 1$ and $\lambda_b = 1$, as done by Stephens (2000 A).

To asses convergence of the samplers, three chains were run for each one, from an over dispersed starting point. The first 300000 iterations were discarded as burn-in and the following 100000 were thinned every 50 iterations to end with a total $T = 2000$ sweeps for each chain. We selected $b = 100$ and evaluated the corresponding diagnostic statistics for each of the resulting 10 batches.

The number of components, k, was used as a "model" indicator and for each example eight observations were selected and the parameter vector $\boldsymbol{\theta}_*$ was formed by all the mean vectors to which each observation was allocated. The first two observations were selected by obtaining the minimum spanning tree and keeping the two observations that were joined by the largest edge in the tree. After removing this edge, we repeated the procedure again for the two resulting data subsets and kept the four observations

that were joined with the largest edge in each tree. Finally, we selected the observations alternating the minimum and maximum with each dimension. We hope to select observations that exhibit a different behaviour across sweeps. That is we expect them to be either between two clusters that compete for the allocation of the observation or potential outliers. The selected observations for the Ruspini data set were: 1, 31, 48, 20, 17, 73, 43 and 44. For the Old Faithful data were: 1, 131, 149, 218, 244, 265, 7 and 89. Results for the Old Faithful and the Ruspini data are shown in Figures ref here and ref here respectively.

In general, we conclude that there is no evidence to support the hypothesis that the chain has not reached equilibrium for the long runs we have used. Trans-dimensional MCMC samplers may require longer runs to reach equilibrium than many MCMC samplers in fixed parameter spaces, particularly in high dimensional problems. Despite having considered a long burn-in period and a long set of iterations, we have seen that the first few hundred iterations still display some instability and could be discarded before carrying out inference.

For the Old Faithful data, the sampler has a mode in an eight-component mixture model with posterior probability, $p(k = 8) = 0.47$. The matrix A below corresponds to the average affinity matrix, some of the values are too small and appear as zeros. The closer pairs of densities are those whose affinity is closer to 1.

$$A = \begin{pmatrix} 0 & 0.0007 & 0.0144 & 0.2549 & 0.0004 & 0.0135 & 0.2326 & 0.0070 \\ 0 & 0 & 0.0005 & 0.0003 & 0.3763 & 0.0339 & 0.0117 & 0.0134 \\ 0 & 0 & 0 & 0.0268 & 0.0034 & 0.0001 & 0.0012 & 0.1815 \\ 0 & 0 & 0 & 0 & 0.0002 & 0.0134 & 0.0234 & 0.0023 \\ 0 & 0 & 0 & 0 & 0 & 0.0040 & 0.0032 & 0.1770 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.2531 & 0.0007 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0076 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

It only makes sense to look at the values of the average affinity matrix and the swap matrix when the stationarity of the chain has been reached. We monitored the entries of the affinity matrix for convenance of presentation the plots are not shown. The initial values showed more variation which was slightly accentuated where there were jumps in dimension. The entries $A(i, j)$ with very small values showed more variability but in general, the matrix shows a stable behavior. We display the matrix A graphically as a dendrogram in Figure 5 (b), the dendrogram is built using $1 - A$ as a dissimilarity matrix in an average linkage hierarchical clustering algorithm.

The results for the swap matrix showed more variability at the beginning of the sampler than those observed for the affinity matrix. However, variation of the swaps when dimension jumps are made shows less effects than with the affinity matrix. The entries with smaller values showed more variability but in general the matrix tends to stabilize after the first 600 monitored iterations. Results are also displayed graphically as a dendrogram that again is built using the dissimilarity matrix $1 - S$, where S is the

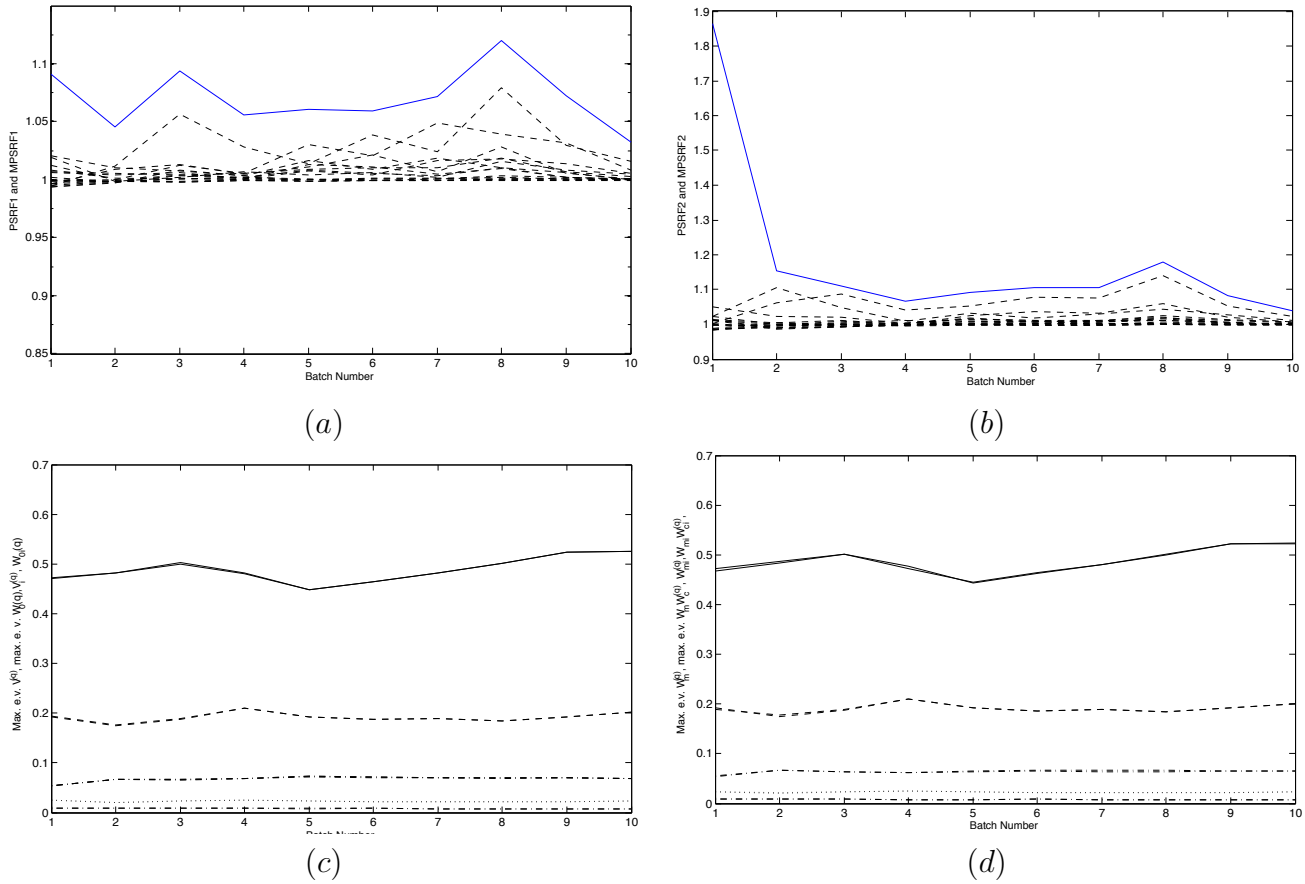


Figure 3: Convergence assessment for Old Faithful data set. *(a)* Solid line: $MPSRF1^{(q)}(\boldsymbol{\theta}_*)$ by batch number q. Dashed lines: $PSRF1^{(q)}(\theta_i)$ by batch number q. *(b)* Solid line: $MPSRF2^{(q)}(\boldsymbol{\theta}_*)$ by batch number q. Dashed lines: $PSRF2^{(q)}(\theta_i)$ by batch number q. *(c)* Solid lines: Maximum eigenvalues for $\widehat{V}^{(q)}(\boldsymbol{\theta}_*)$ and $Wc^{(q)}(\boldsymbol{\theta}_*)$ by batch number q. Dashed lines: $\widehat{V}^{(q)}(\theta_i)$ and $Wc^{(q)}(\theta_i)$ (for some observations) by batch number q. *(d)* Solid lines: Maximum eigenvalues for $Wm^{(q)}(\boldsymbol{\theta}_*)$ and $WmWc^{(q)}(\boldsymbol{\theta}_*)$ by batch number q. Dashed lines: $Wm^{(q)}(\theta_i)$ and $WmWc^{(q)}(\theta_i)$ (for some observations) by batch number q.

swap matrix, in Figure 5 (c).

$$S = \begin{pmatrix} 0 & 0.0006 & 0.0004 & 0.1086 & 0.0001 & 0.0055 & 0.0959 & 0.0015 \\ 0 & 0 & 0.0005 & 0 & 0.1745 & 0.0225 & 0.0049 & 0.0021 \\ 0 & 0 & 0 & 0.0051 & 0.0004 & 0 & 0.0001 & 0.0840 \\ 0 & 0 & 0 & 0 & 0 & 0.0121 & 0.0110 & 0.0005 \\ 0 & 0 & 0 & 0 & 0 & 0.0005 & 0.0014 & 0.0621 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.1080 & 0.0002 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0041 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

From the dendrograms in Figures 5 (b) and (c) we would merge the components into two groups, the first including components 4, 1, 7 and 6 and the second group includes

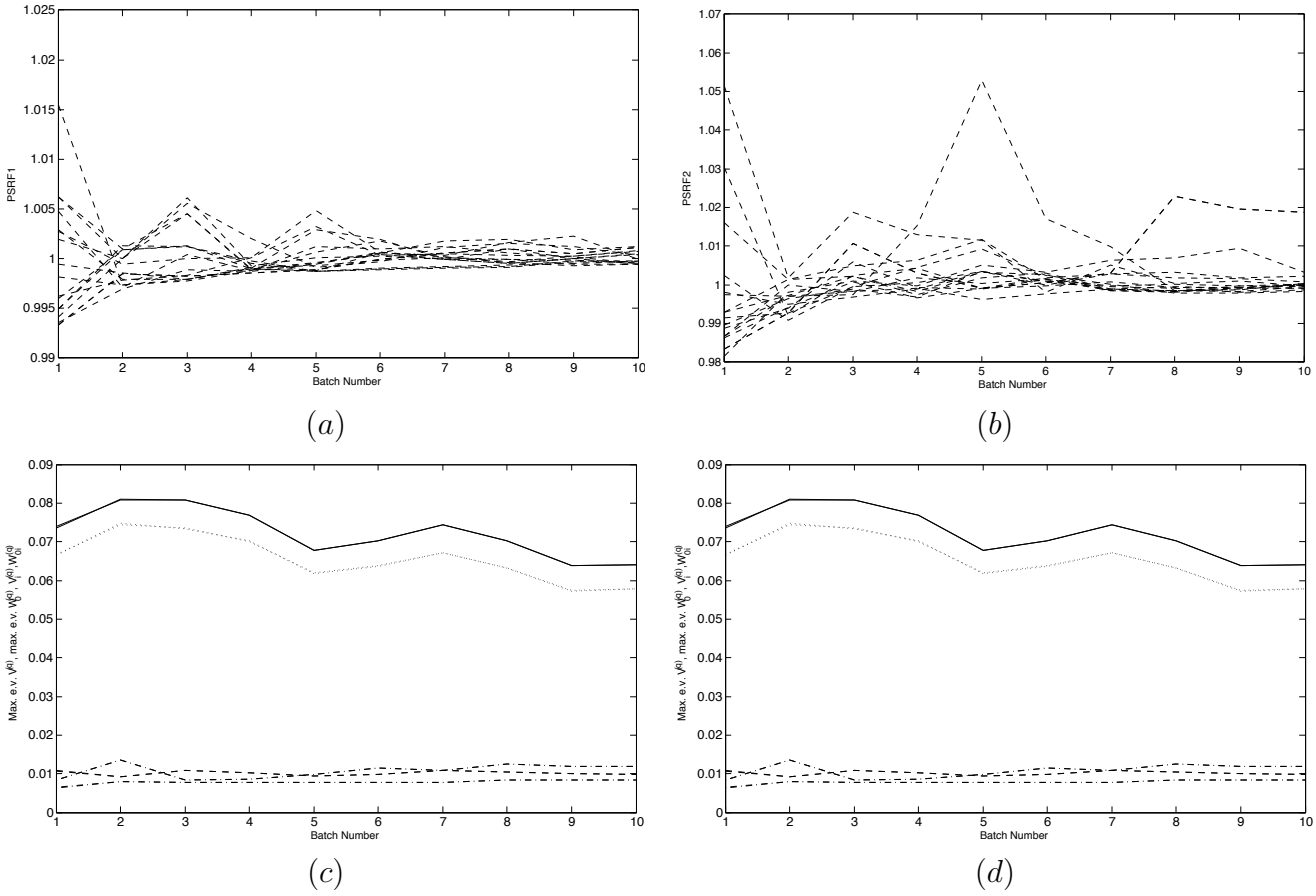


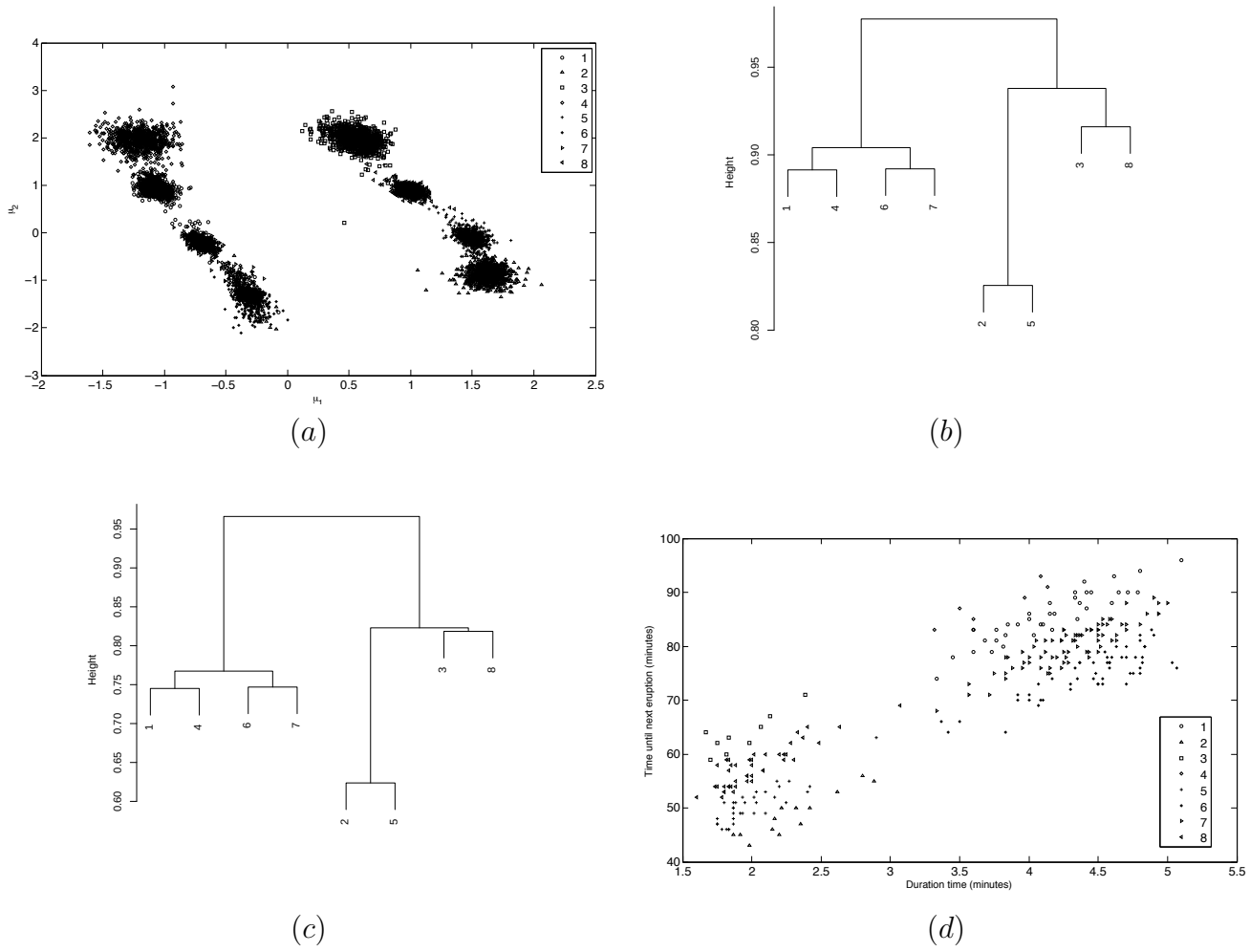
Figure 4: Convergence assessment for Ruspini data set. *(a)* Dashed lines: $PSRF1^{(q)}(\theta_i)$ by batch number q. *(b)* Dashed lines: $PSRF2^{(q)}(\theta_i)$ by batch number q. *(c)* Solid lines: Maximum eigenvalues for $\widehat{V}^{(q)}(\boldsymbol{\theta}_*)$ and $Wc^{(q)}(\boldsymbol{\theta}_*)$ by batch number q. Dashed lines: $\widehat{V}^{(q)}(\theta_i)$ and $Wc^{(q)}(\theta_i)$ (for some observations) by batch number q. *(d)* Solid lines: Maximum eigenvalues for $Wm^{(q)}(\boldsymbol{\theta}_*)$ and $WmWc^{(q)}(\boldsymbol{\theta}_*)$ by batch number q. Dashed lines: $Wm^{(q)}(\theta_i)$ and $WmWc^{(q)}(\theta_i)$ (for some observations) by batch number q.

components 3, 8, 5 and 2. The resulting classification has some slight variations from the results obtained from the nonrestricted model but in general it is consistent with what one would visually expect.

Results for the Ruspini data set show the mode is at a 4-component mixture with posterior probability, $p(k = 4) = 0.94$. The corresponding mean affinity and swap matrices are:

$$A = \begin{pmatrix} 0 & 0.0006 & 0.0164 & 0.0068 \\ 0 & 0 & 0.0180 & 0.0103 \\ 0 & 0 & 0 & 0.0001 \\ 0 & 0 & 0 & 0 \end{pmatrix}; \quad S = \begin{pmatrix} 0 & 0 & 0.0016 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

The results are shown in Figures 6 (a)-(d) and from the affinity and swap matrices we would not merge any components. This analysis will define four groups in this data set and the corresponding classification is given in Figure 6 (d). In this case the rescaling



(a) (b) (c) (d)

Figure 5: Old Faithful data. *(a)* Sampled mean values for a 7-component mixture after removing the label switching, for the rescaled data set. *(b)* Swap matrix. *(c)* Affinity matrix. *(d)* Classification of the Old Faithful data.

of the data gives a smaller number of components but the classification is consistent with the classification obtained with the nonrestricted and restricted models fitted to the original data.

6 Concluding remarks

We begin with some remarks about the performance of multivariate normal distributions when faced with data that is inconsistent with such a model. We have often observed that small weighted components are often included possibly to accommodate departures from normality, including outliers. The fitted model frequently copes with non-normality by preferring a small number of highly dispersed components to a more complex model with a larger number of components.

The intuitive idea we pursued in this paper was to exclude the assumption that one component is used to describe one cluster in a model-based context. Instead, we allowed

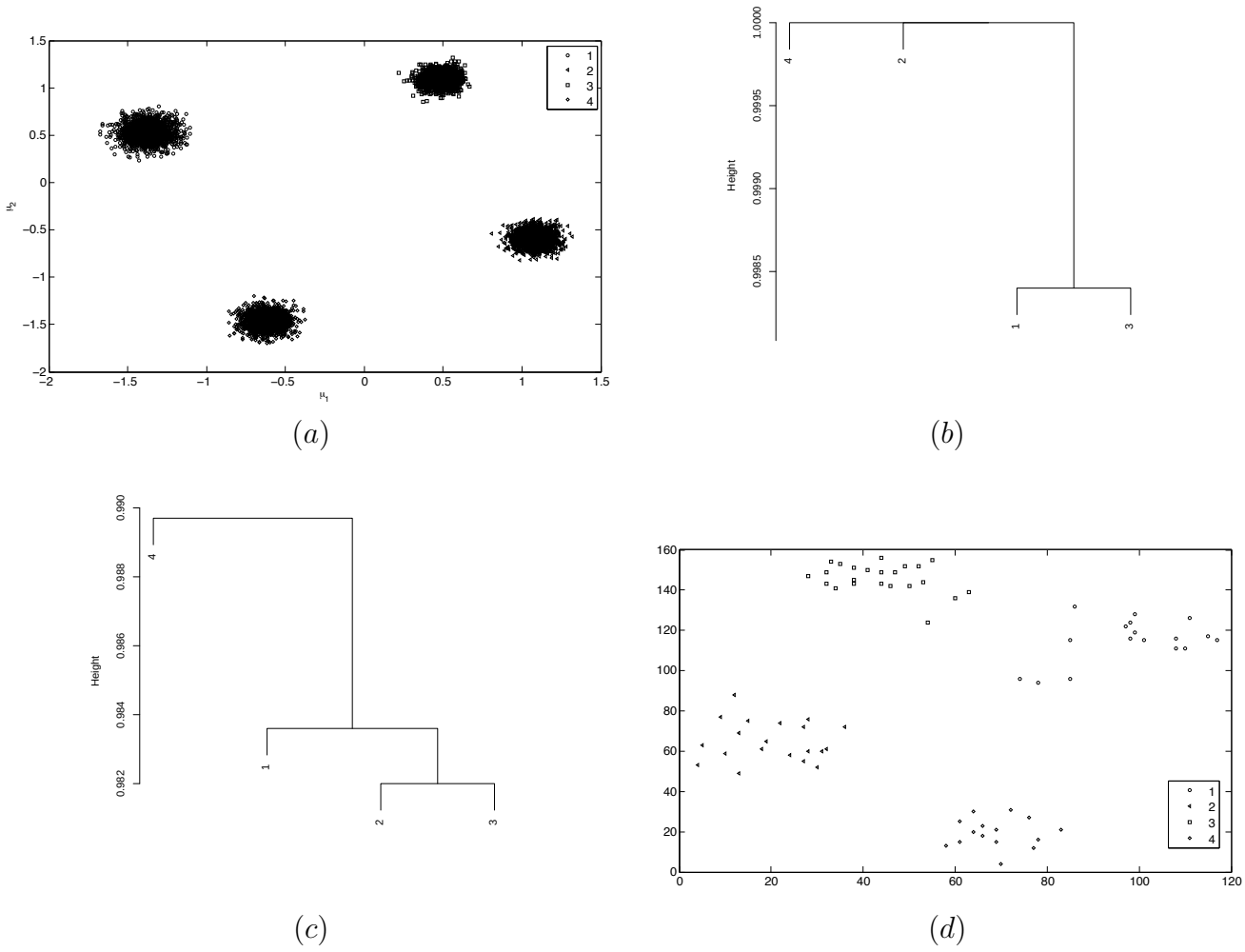


Figure 6: Ruspini data. *(a)* Sampled mean values for a 4-component mixture after removing the label switching for the rescaled data set. *(b)* Swap matrix. *(c)* Affinity matrix. *(d)* Classification of the Ruspini data.

a submixture of such components to represent a cluster. Thus we need to produce more, possibly many more model components than groups and subsequently find a method of combining some components into submixtures that describe sensible clusters. This can be achieved by restricting the covariance structure of the multivariate normal components which also has the benefit of introducing simpler parametric forms for the basic element of our model, allowing for a simpler fitting. More flexibility can easily be introduced by relaxing the assumptions made for the covariance matrices used in this paper. However, care is needed to avoid the extreme situation where one component is used to describe a single observation which belongs to a compact group.

In cluster analysis based on submixture models, the most significant aspect to consider was the proportion of observations that were exchanged between components, always verifying that those components are closer to each other than other pairs in the fitted model. That will determine which of the pairs should be merged helping to identify some overlapping clusters in a more efficient way. It also would avoid merging pairs

of components which have swapped observations mainly when the sampled values for the parameters of the components are in the tails of the corresponding distributions. However, we will always have the option of combining two different groups that are very close to each other and have many observations in the *boundaries* of the two components. In this case, an important number of observations would be exchanged between these components and the criteria would suggest to consider them as only one group.

Acknowledgments

The author would like to thank the Collegio Carlo Alberto, Turin, Italy. She has written this paper while enjoying a Visiting Fellowship awarded by this institution within the "de Castro" Statistics Initiative.

Appendix

The Bhattacharyya (1943) distance also known as the affinity between two distributions P_1 and P_2, with corresponding densities p_1, p_2 is given as

$$A(p_1, p_2) = \int_{\mathbf{y}} \sqrt{p_1}\sqrt{p_2}\, d\mathbf{y}. \tag{6.1}$$

We calculate the affinity between two multivariate normal distributions, $N(\mathbf{y}|\boldsymbol{\mu}_1, \Sigma_1)$ and $N(\mathbf{y}|\boldsymbol{\mu}_2, \Sigma_2)$.

$$\begin{aligned} A(p_1, p_2) &= \int_{\mathbf{y}} \left[(2\pi)^{-\frac{p}{2}}|\Sigma_1|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_1)^T\Sigma_1^{-1}(\mathbf{y}-\boldsymbol{\mu}_1)\right\}\right]^{\frac{1}{2}} \times \\ &\times \left[(2\pi)^{-\frac{p}{2}}|\Sigma_2|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_2)^T\Sigma_1^{-1}(\mathbf{y}-\boldsymbol{\mu}_2)\right\}\right]^{\frac{1}{2}} d\mathbf{y} \\ &= \int_{\mathbf{y}} (2\pi)^{-\frac{p}{2}}|\Sigma_1|^{-\frac{1}{4}}|\Sigma_2|^{-\frac{1}{4}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_1)^T(2\Sigma_1)^{-1}(\mathbf{y}-\boldsymbol{\mu}_1)\right\} \times \\ &\times \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_2)^T(2\Sigma_2)^{-1}(\mathbf{y}-\boldsymbol{\mu}_2)\right\} d\mathbf{y}. \end{aligned} \tag{6.2}$$

Recall the following property of Gaussian densities: the product of two Gaussian densities is proportional to another Gaussian density

$$N(\mathbf{y}|\boldsymbol{\mu}_1, \Sigma_1) \times N(\mathbf{y}|\boldsymbol{\mu}_2, \Sigma_2) \propto N(\mathbf{y}|\boldsymbol{\mu}_*, \Sigma_*), \tag{6.3}$$

where $\boldsymbol{\mu}_* = \Sigma_*(\Sigma_1^{-1}\boldsymbol{\mu}_1 + \Sigma_2^{-1}\boldsymbol{\mu}_2)$ and $\Sigma_* = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}$. Rewriting A in expression 6.2 to be able to use the latter property we get

$$\begin{aligned} A(p_1, p_2) &= (2\pi)^{\frac{p}{2}}|\Sigma_1|^{-\frac{1}{4}}|\Sigma_2|^{-\frac{1}{4}}|2\Sigma_1|^{\frac{1}{2}}|2\Sigma_2|^{\frac{1}{2}} \times \\ &\times \int_{\mathbf{y}} (2\pi)^{-\frac{p}{2}}|2\Sigma_1|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_1)^T(2\Sigma_1)^{-1}(\mathbf{y}-\boldsymbol{\mu}_1)\right\} \times \\ &\times (2\pi)^{-\frac{p}{2}}|2\Sigma_2|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_2)^T(2\Sigma_2)^{-1}(\mathbf{y}-\boldsymbol{\mu}_2)\right\} d\mathbf{y}. \end{aligned} \tag{6.4}$$

Rewriting expression 6.4 in terms of a normal distribution with mean vector and covariance matrix

$$\begin{aligned}\mu_* &= \Sigma_*((2\Sigma_1)^{-1}\boldsymbol{\mu}_1 + (2\Sigma_2)^{-1}\boldsymbol{\mu}_2),\\ \Sigma_* &= ((2\Sigma_1)^{-1} + (2\Sigma_2)^{-1})^{-1},\end{aligned}$$

and adding the required normalizing constant, we get

$$\begin{aligned}A(p_1, p_2) &= |\Sigma_1|^{-\frac{1}{4}}|\Sigma_2|^{-\frac{1}{4}}|\Sigma_*|^{\frac{1}{2}} \times \\ &\quad \times \exp\left\{-\frac{1}{4}\left(\boldsymbol{\mu}_1^T\Sigma_2^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T\Sigma_2^{-1}\boldsymbol{\mu}_2\right) + \frac{1}{2}\left(\boldsymbol{\mu}_*^T\Sigma_*\boldsymbol{\mu}_*\right)\right\},\end{aligned}$$

where now $\boldsymbol{\mu}_* = ((2\Sigma_1)^{-1}\boldsymbol{\mu}_1 + (2\Sigma_2)^{-1}\boldsymbol{\mu}_2)$.

In the particular case where Σ_1 and Σ_2 are diagonal matrices, the affinity between p_1 and p_2 is given by

$$A(p_1, p_2) = \prod_{i=1}^{p}\left(\frac{2\sigma_{1,i}\sigma_{2,i}}{\sigma_{1,i}^2 + \sigma_{2,i}^2}\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{4}\sum_{i=1}^{p}\frac{(\boldsymbol{\mu}_{1,i} - \boldsymbol{\mu}_{2,i})^2}{\sigma_{1,i}^2 + \sigma_{2,i}^2}\right\}.$$

References

Banfield, D. B. and Raftery, A. E. (1993). Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, **49**, 803–821.

Bhattacharyya, A (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, **35**, 99 – 110.

Bensmail, H. and Celeux, G. and Raftery, A. E. and Robert, C. P. (2003). Inference in model-based cluster analysis. *Statatistics and Computing*, **7**, 1–10.

Brooks, S. P. and Giudici, P. and Phillipe, A. (2003). Nonparametric convergence assessment for MCMC model selection. *Journal of Computational and Graphical Statistics*, **12**, 1–22.

Castelloe, J.M. and Zimmerman, D.L. Convergence assessment for reversible jump MCMC samplers *Technical report 313. Department of Actuarial Sciences, University of Iowa.*

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern recongnition*, **28**, 781–793.

Celeux, G. and Hurn, M. and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, **95**, 957–970.

Dellaportas, P. and Papageorgiou, I. (2006). Multivariate mixtures of normals with an unknown number of components. *Statistics and Computing*, **16**, 57–68.

Fuentes-García, R. S. (2004). *Bayesian model-based cluster analysis.* Ph.D. Thesis University of Bath.

Fritsch, A. and Ickstadt, I. (2009). Improved Criteria for Clustering Based on the Posterior Similarity Matrix. *Bayesian Analysis*, **4**, 367–392.

Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, **96**, 194–209.

Green, P. J. (1995). Reversible Jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Härdle, W. (1991). *Smoothing techniques with implementation in S.* Springer, New York.

Hurn, M., Justel, A. and Robert, C. P. (2003). Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics*, **12** 55–79.

Jasra A., Holmes, C.C. and Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian Mixture Modelling. *Statistical Sciences*, **20**, 56–67.

Lau, J.W. and Green, P. J. (2003). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, **16** 526–558.

Lijoi, A. and Mena, R. H. and Prünster, I.(2007). Controlling the reinforcement in Bayesian nonparametric mixture models. *Journal of the Royal Statistical Society B*, **69**, 715–740.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* Wiley Series in Probability and Statistics, John Wiley & Sons.

Richardson, S. and Green, P. J. (1997). On Bayesian Analysis of Mixtures with an unknown Number of Components. *Journal of the Royal Statistical Society: Series B*, **59**, 731–792.

Ripley, B. D. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society: Series B*, **39**, 172–212.

Ruspini, E. H. A new approach to clustering. *Information and Control*, **15**, 22–32.

Sisson, S. A. (2005). Trans-dimensional Markov chains: A decade of progress and future perspectives. *Journal of the American Statistical Association*, **100**, 1077–1089.

Sisson, S. A. and Y. Fan A distance-based diagnostic for trans-dimensional Markov chains. *Journal of Statistics and Computing*, **17**, 357–367.

Stephens, M. (2000 A). Bayesian Analysis of mixture Models with an unknown Number of Components- an alternative to Reversible Jump Methods. *Annals of Statistics*, **28**, 40–74.

Stephens, M. (2000 B). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B*, **62**, 795–809.

Venables, W. N. and Ripley, B. D. (1997 2nd Ed.). *Modern Applied Statistics with S-plus*, Springer, New York. 2nd Edition.

Zhihua, Z. and Kapluk, C. and Yiming, W. and Chibiao, C. (2004). Learning a multivariate Gaussian mixture model with reversible jump MCMC algorithm. *Statistics and Computing*, **14**, 343–355.